A Novel Approach Based on Deep Learning for Violence Detection in Public Places

Shah Faisal Khan[1], Said Khalid Shah[1], Faheem Ullah Khan[2], Wasiat Khan[2], Fouzia Idrees[3]
[1]Department of Computer Science, University of Science and Technology Bannu, KP, Pakistan
[2]Department of Software Engineering, University of Science and Technology Bannu, KP, Pakistan
[3]Department of Computer Science, Shaheed Benazir Bhutto Women University Peshawar, KP, Pakistan
*Correspondence: faisalabidi3115@gmail.com, skhalids7575@gmail.com

Citation | Khan. S. F, Shah. S. K, Khan. F. U, Khan. W, Idrees. F, "A Novel Approach Based on Deep Learning for Violence Detection in Public Places", IJIST, Vol. 08 Issue. 01 pp 406-415, February 2026
Received | January 07, 2026 **Revised |** February 07, 2026 **Accepted |** February 10, 2026 **Published |** February 14, 2026.

This study presents a lightweight deep learning pipeline for violence detection in public place surveillance imagery. The central idea is to perform offline key frame selection before model training so that redundant, transitional, and visually ambiguous frames are discarded, and only informative violent or non-violent scenes are retained. Unlike clip-level methods that rely on dense temporal stacks or computationally heavy 3D networks, the proposed approach uses curated key frames to construct a balanced 7,000-image dataset and then trains a transfer learning classifier on those frames. Images are resized to 256 × 256, normalized, and augmented before being processed by a ResNet-50 backbone, followed by two fully connected layers of 512 and 256 units and dropout regularization. On Hockey Fight dataset test images, the model achieves 97.25% accuracy, 97.62% precision, 96.45% recall, and 97.03% F1-score, outperforming EfficientNet-B0 and VGG-19 under the same experimental pipeline. The results indicate that careful offline frame curation can provide a practical compromise between computational efficiency and recognition performance for early violence screening in surveillance settings. The paper also discusses current limitations, including manual curation bias, the absence of explicit temporal modeling, and the lack of actor-level localization.

Keywords: Violence Detection; Public-Place Surveillance; Key-Frame Selection; Activity Recognition; Transfer Learning; ResNet-50.

Introduction:

Video surveillance has become a standard component of public-safety infrastructure in smart cities, transport hubs, campuses, and commercial environments. The rapid growth of visual monitoring systems has increased the demand for automatic recognition models that can assist human operators by flagging unusual or dangerous events in real time [1][2]. Among these tasks, violence detection is especially important because delayed intervention can lead to injuries, property damage, and broader public panic. A practical surveillance model must therefore balance recognition accuracy, computational efficiency, and robustness to the variability of real-world scenes.

Despite substantial progress in deep learning, violence detection remains difficult in unconstrained public environments. Many existing systems are built for clip-level classification and depend on optical flow, stacked video frames, recurrent networks, or other spatiotemporal modules that increase computational cost [3][4][5][6][7][8]. In addition, several published methods focus primarily on whether a video clip is violent, while giving less attention to where the violent interaction occurs, how multiple people are socially engaged in the event, or how redundant and ambiguous frames affect model learning. For public-place monitoring, long video streams contain large numbers of repetitive frames, and training directly on all of them can introduce redundancy, blur, transition artifacts, and label noise.

This study investigates a lighter alternative: instead of relying on dense temporal input, it first performs offline key-frame selection to construct a balanced and visually informative image dataset, and then trains a transfer learning classifier on those curated frames. The proposed contribution is therefore not a full spatiotemporal localization model; rather, it is a scene-level violence detector designed for early screening. Its novelty lies in using key-frame selection as a dataset-construction stage that removes redundant and uncertain frames before training, thereby improving data quality for a lightweight classifier. The revised manuscript also positions this contribution more explicitly against existing literature, clarifies the experimental setup, and discusses the limitations that remain for temporal reasoning, localization, and annotation reliability.

The main contributions of the study are as follows:

An offline key-frame curation pipeline that constructs a balanced 7,000-image dataset by filtering visually ambiguous and redundant frames from public online surveillance sources.

A transfer learning framework based on a pre-trained ResNet-50 backbone with a lightweight two-layer classification head for binary violence recognition.

A unified experimental comparison against alternative pre-trained backbones using the same data pipeline, training environment, and evaluation metrics.

A critical discussion of methodological limitations, including manual curation bias, incomplete temporal modeling, and the absence of actor-level localization.

Related Work:

Early violence detection research relied on hand-crafted motion descriptors, optical-flow statistics, or engineered scene cues. Representative studies by [9], [10], and [11] established the importance of motion and interaction patterns in surveillance video, but these methods generally required carefully designed features and struggled with large appearance variation.

With the rise of deep learning, violence detection systems began to use CNNs, 3D CNNs, and hybrid CNN–SVM or CNN–RNN architectures. [4] showed that 3D convolutions can capture spatiotemporal patterns effectively, while [5] combined 3D CNNs with SVM classification. [6] proposed an efficient surveillance-oriented framework, and [12] used a deep learning pipeline for violence recognition in video data. These studies improved accuracy, but most of them still depend on video clips or motion modeling during inference.

Recent work has expanded this direction using transformers, skeleton cues, attention modules, and increasingly specialized datasets. [7] combined human skeletons, change detection, and ConvLSTM aggregation to obtain an efficient real-time system. [8] proposed CrimeNet, a Vision Transformer with neural structured learning that improves ROC-AUC and reduces false positives across challenging datasets. [13] employed key-frame identification with 3D CNNs and CBAM attention, [14] introduced a shallow 3D CNN for more lightweight inference, and [15] used a video Swin Transformer to analyze crowd size and violence level. Related work on interaction recognition also shows that modeling interpersonal behavior can be highly informative for surveillance understanding [16] Beyond these studies, 2024 literature further broadened the field with attention-guided 2D CNNs [17], end-to-end real-time 2D CNN pipelines [18], fused-feature surveillance models [19], comprehensive reviews of research gaps and cross-dataset weakness [20][21], transformer-based post-processing for cross-dataset robustness [22], new benchmark datasets [23], workshop-scale transformer systems such as CUE-Net [24], multi-head attention plus LSTM pipelines [25], and efficient deep-feature sequential analysis for more realistic settings [26].

Although these studies are strong, they also reveal a gap relevant to the present paper. Most recent methods are clip-based and therefore require temporal stacks, pose extraction, graph reasoning, or transformer-style video modeling [17][18][22][24][25][26]. Such designs are powerful but can be computationally expensive and difficult to deploy in settings where quick screening is needed. The 2025 literature continues this trend through stronger image-based baselines [27], hybrid surveillance systems that combine violence recognition with anomaly detection and re-identification [28], graph-based skeleton analysis [29], identity-aware graph plus 3D-CNN fusion [30], and keyframe-focused excitation networks [31]. At the same time, recent reviews and benchmark papers emphasize persistent problems in generalization, annotation consistency, and realistic public-surveillance coverage [20][21][22][23]. The present work addresses only part of this gap by improving frame quality through offline key-frame selection and by preserving visible interaction context inside each training image. However, it remains a scene-level classifier and does not solve full actor-level localization or temporal reasoning.

Proposed Method:

Figure 1 summarizes the overall workflow adopted in this study. The pipeline begins with frame collection from public online sources, proceeds through offline key-frame selection, preprocessing, and augmentation, and ends with model training, testing, and scene-level classification.

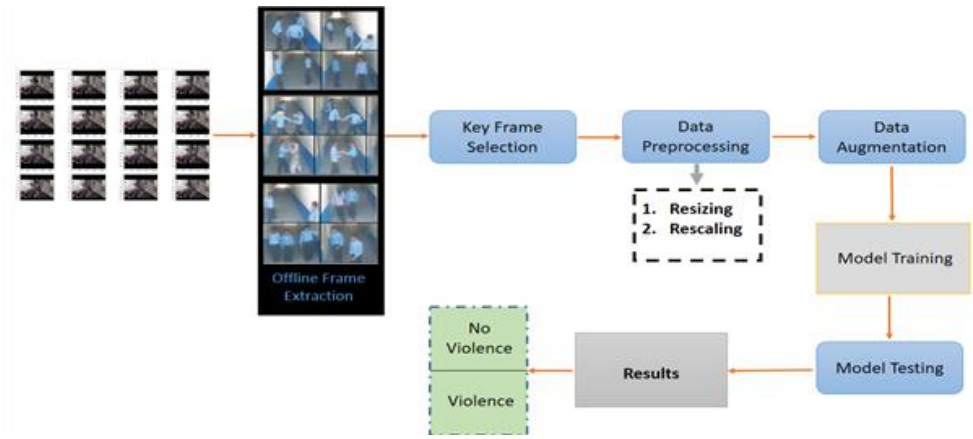

Figure 1. Overall workflow of the proposed violence-detection pipeline.

Dataset Construction and Offline Key-Frame Selection:

The training data were assembled from public online violence and non-violent sources. The primary candidate pool reported in the original study contained 11,063 frames, including

5,832 violent frames and 5,231 non-violent frames. Additional samples were collected from Rob flow Universe search results related to violence-related imagery. After screening and class balancing, the final curated dataset contained 7,000 images, with 3,500 violent and 3,500 non-violent samples.

The key novelty of the dataset pipeline is the role assigned to key-frame selection. In many prior studies, key-frame extraction is part of a video model and is used to select frames dynamically during temporal inference [13] In contrast, the present work uses key-frame selection offline, before training, as a data-curation stage. The purpose is to remove redundant, transitional, and visually uncertain frames so that the learning algorithm sees a more informative set of examples. This makes the selection step model-agnostic and reduces the amount of low-value visual redundancy passed to the classifier.

Because the curation process was manual, it can introduce selection bias toward visually obvious cases. The curated dataset should therefore be interpreted as a high-confidence subset rather than a statistically exhaustive sample of all violence scenarios in public places. The revised manuscript explicitly acknowledges this limitation. A stronger protocol for future work would use multiple annotators, documented annotation guidelines, and inter-rater agreement statistics.

Table 1. Dataset sources and construction of the final curated image set.

Source	Role in pipeline	Reported raw material	Output after curation
Public online violence/non-violence repository	Primary candidate pool	11,063 frames (5,832 violent; 5,231 non-violent)	Visually screened to retain clear violent/non-violent interactions
Rob flow Universe search results	Supplementary samples	Additional violence-related images/frames	Screened with the same key-frame selection criteria
Final curated dataset	Model development	Balanced subset	7,000 images (3,500 violent; 3,500 non-violent)

Preprocessing and Augmentation:

All input images were resized to 256 × 256 pixels and rescaled to the [0, 1] range before model training. This standardization ensured that the transfer learning backbone received a uniform input size and normalized intensity range. The dataset remained class-balanced after curation, which helped reduce bias between violent and non-violent classes.

To improve generalization, the training pipeline used data augmentation through image flipping, rotation, shift, zoom, shear, and brightness variation. These operations increase appearance diversity and help the model adapt to changes in viewpoint, motion intensity, and illumination. The archived experimental notes list the augmentation types but do not preserve their exact numeric ranges; this is acknowledged later as a reproducibility limitation.

Model Architecture and Transfer Learning Strategy:

The proposed classifier uses a pre-trained ResNet-50 backbone as a generic feature extractor and appends a lightweight binary classification head. After feature extraction and pooling, two fully connected layers with 512 and 256 units are used, and each dense layer is followed by dropout with a rate of 0.3. The final output layer performs violent vs. non-violent scene classification.

Figure 2 provides a clearer summary of the architecture, including the standard feature-map sizes produced by ResNet-50 for a 256 × 256 input. Because ResNet-50 is a mature backbone with well-understood stage-wise dimensionality, the feature-map sizes can be reported explicitly even when the classifier head is task-specific.

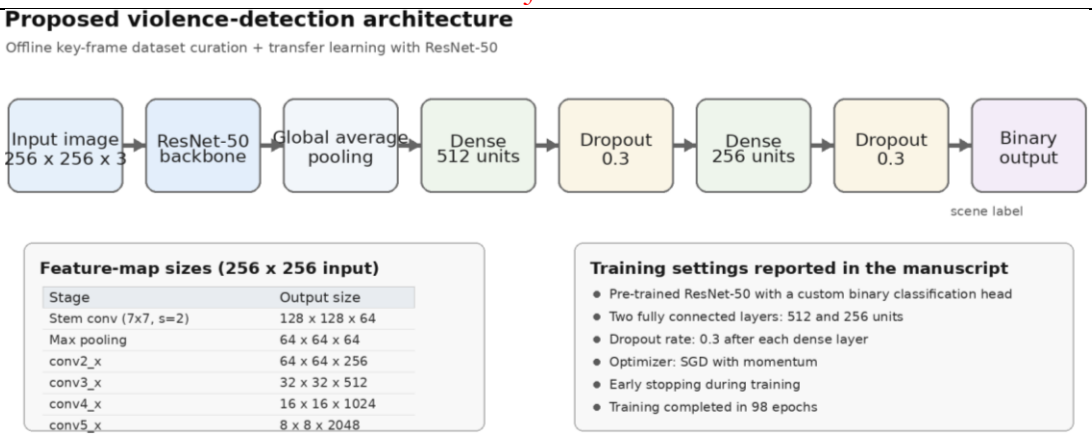

Figure 2. Revised architecture diagram of the proposed ResNet-50-based model.

Training Configuration and Evaluation Protocol:

Training was conducted in a Google Colab environment using an NVIDIA T4 GPU with 12 GB of RAM. The revised manuscript preserves the settings reported in the archived study: transfer learning with ResNet-50, the two-layer classifier head, dropout regularization, stochastic gradient descent with momentum, and early stopping. The training process ran for 98 epochs. A held-out validation set was used during model monitoring, as reflected by the training and validation curves.

To keep the backbone comparison fair, EfficientNet-B0, VGG-19, and ResNet-50 were compared under the same preprocessing, augmentation, hardware environment, optimization setup, and evaluation metrics. Only the backbone feature extractor differed across these experiments.

The available evaluation metrics are accuracy, precision, recall, and F1-score. The present revision also adds extra visual summaries of model performance. However, confusion matrices and ROC curves cannot be reconstructed from the archived manuscript alone because per-sample prediction scores were not retained in the supplied material.

Table 2. Reported training settings and reproducibility notes.

Item	Reported setting
Training environment	Google Colab with NVIDIA T4 GPU (12 GB RAM)
Input size	256 × 256 RGB
Normalization	Pixel values rescaled to [0, 1]
Backbone	Pre-trained ResNet-50
Classifier head	Dense(512) → Dropout(0.3) → Dense(256) → Dropout(0.3) → Binary output
Optimizer	SGD with momentum
Stopping rule	Early stopping based on held-out validation performance
Epochs	98
Metrics	Accuracy, precision, recall, and F1-score
Reproducibility notes	Exact batch size, split ratios, interpolation kernel, learning-rate schedule, and layer-freezing depth were not preserved in the archived notes.

Results and Discussion:

Training Behavior:

Figure 3 shows the recorded training and validation curves. The training accuracy rises rapidly in the early epochs and then converges gradually, while the validation curve remains stable with a relatively small gap from the training curve. The loss curves follow the same

pattern: both training and validation loss decrease steadily, which indicates stable optimization and only moderate overfitting.

These curves support the decision to use transfer learning and dropout regularization. For a curated image dataset of this size, the pre-trained backbone provides a strong initialization and helps the model reach high performance without the significantly larger computational cost of training a deep model from scratch.

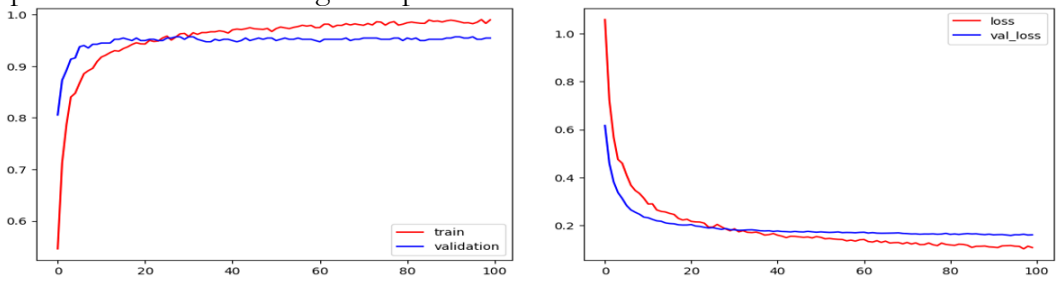

Figure 3. Training and validation accuracy/loss curves are reported in the study.

Test-Set Performance:

Table 3 and Figure 4 summarize the aggregate performance of the proposed model on Hockey Fight dataset test images. The highest value is obtained for precision (97.62%), while recall remains strong at 96.45%, indicating that the model retrieves most violent scenes while keeping false positives low.

Table 3. Evaluation results of the proposed model on test images.

S. No.	Metric	Value (%)
1	Accuracy	97.25
2	Precision	97.62
3	Recall	96.45
4	F1-score	97.03

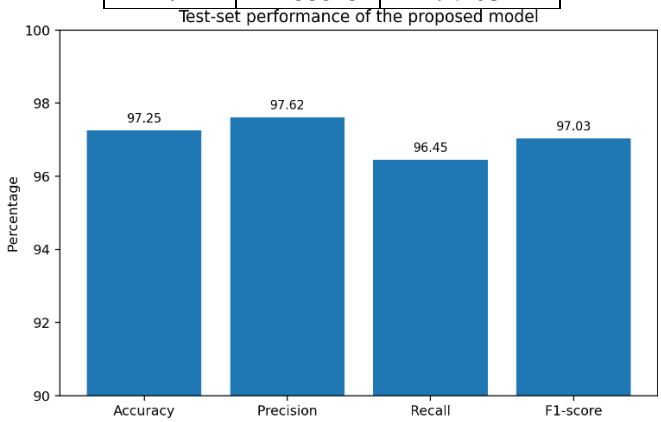

Figure 4. Additional visualization of the model's aggregate test-set metrics.

Comparison with Other Pre-Trained Backbones:

To evaluate the suitability of the chosen backbone, the same training pipeline was run with EfficientNet-B0, VGG-19, and ResNet-50. Table 4 and Figure 5 show that ResNet-50 delivers the strongest overall results across all reported metrics. These results suggest that the proposed classifier head benefits from the representational depth and stability of ResNet-50 while remaining lighter than many full video-based models.

The comparison is fair in the sense that the data preparation, augmentation policy, optimization setup, and evaluation metrics were held constant. Only the feature extractor was changed.

Table 4. Comparison of pre-trained backbones under the same experimental pipeline.

Backbone	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
EfficientNet-B0	96.36	97.05	95.81	96.24
VGG-19	95.10	96.55	94.42	95.83
ResNet-50	97.25	97.62	96.45	97.03

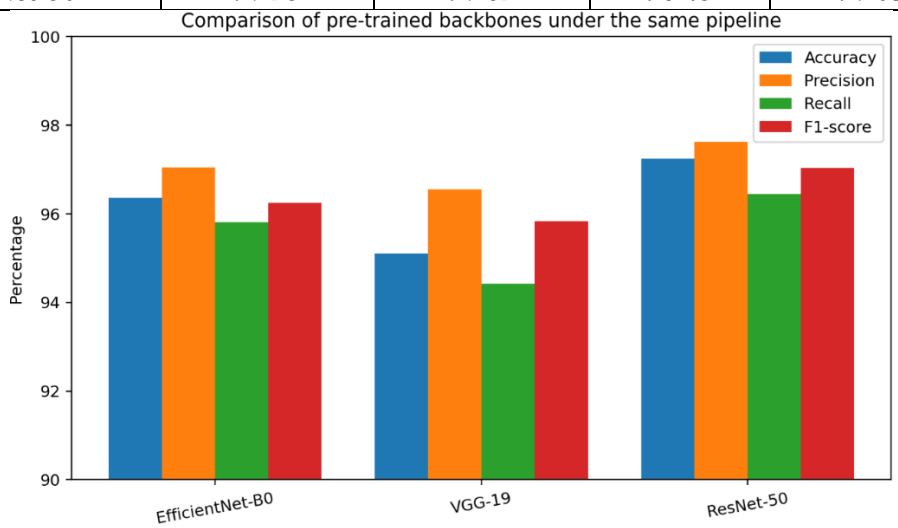

Figure 5. Additional comparison plot for the evaluated pre-trained backbones.

Comparison with Representative Baselines on the Same Dataset:

Cross-dataset comparison can be misleading because violence datasets differ substantially in scene structure, camera angle, crowd density, and annotation policy. For that reason, Table 5 compares the proposed model only with representative baselines reported on the Hockey Fight dataset. Under the same benchmark setting, the proposed ResNet-50-based system achieves the highest accuracy in the comparison set.

Table 5. Comparison with representative baselines evaluated on the same Hockey Fight benchmark.

Representative baseline	Dataset	Accuracy (%)
CNN + SVM	Hockey Fight	94.50
Hough Forests + CNN	Hockey Fight	94.60
VGG-16	Hockey Fight	89.10
VGG-19 + LSTM	Hockey Fight	96.33
Proposed ResNet-50 + dense head	Hockey Fight	97.25

Discussion with Recent Literature (2024-2025):

The proposed model performs well because the curated training set reduces redundancy and because the pre-trained backbone transfers strong generic visual features to the violence-detection task. This lightweight design contrasts with many recent 2024-2025 approaches that depend on explicit temporal modeling, attention modules, graph reasoning, or transformer backbones [17][18][22][24][25][26][29][30][31]. In comparison with those methods, the present framework is simpler to train and more computationally economical because it avoids clip stacks, optical flow branches, pose graphs, and heavy video transformers. For practical surveillance screening, this simplicity is a meaningful advantage, especially where hardware or latency constraints make full video modeling difficult.

A comparison with recent literature also helps position the reported 97.25% accuracy more carefully. Recent methods, such as the attention-based 2D CNN model of [17] the real-time 2D CNN framework of [18], and the fused-feature F3DNN-Net model [19], confirm that strong violence recognition can be achieved without always resorting to the heaviest architectures. At the same time, newer systems such as CUE-Net [24], graph-based skeleton

modeling, and IDG-ViolenceNet show that explicit interaction and temporal reasoning can push performance further on challenging datasets, although at the cost of additional model complexity. The present results, therefore, support a balanced interpretation: careful offline key-frame curation can make a scene-level ResNet-50 pipeline highly competitive on Hockey Fight, but richer temporal and relational representations are still likely to be necessary for harder public-space scenarios and broader cross-dataset transfer.

Reproducibility and Limitations:

At the same time, the approach has clear limitations. First, the model operates on individual frames and therefore cannot explicitly reason over motion trajectories, action duration, or temporal escalation. Second, it predicts scene-level violence rather than spatially localizing the actors or subgroups responsible for the event. This matters because public-place violence is often social and relational, involving both aggressors and bystanders in the same field of view.

Third, the manual frame-selection stage may bias the dataset toward visually obvious cases and may underrepresent borderline or ambiguous situations. Formal annotation-reliability measures were not included in the archived study. Finally, although the revised manuscript adds extra plots, confusion matrices, and ROC curves still require per-sample prediction outputs that were not preserved in the supplied material. Future experimental releases should retain those outputs and report exact split ratios, batch size, learning-rate scheduling, interpolation mode, and layer-freezing details.

Conclusion and Future Work:

This paper presented a revised and more clearly documented violence-detection framework for public places based on offline key-frame selection and transfer learning. The method constructs a balanced 7,000-image dataset from public online sources, removes visually uncertain and redundant frames before training, and uses a ResNet-50 backbone with a lightweight dense classification head. Under the reported evaluation setting, the model achieves 97.25% accuracy, 97.62% precision, 96.45% recall, and 97.03% F1-score on Hockey Fight dataset test images and outperforms the other pre-trained backbones examined in the same pipeline.

Scientifically, the study contributes a dataset-centric perspective to violence detection: instead of increasing model complexity, it improves data quality through offline curation. However, the approach remains limited by manual selection bias, incomplete temporal modeling, and the absence of actor-level localization. Future work should therefore combine curated key-frame selection with stronger temporal encoders, explicit localization of violent subgroups, more transparent multi-annotator labeling, and richer evaluation artifacts such as confusion matrices, ROC curves, and cross-dataset generalization studies.

References:

[1] L. Minh Dang, Kyungbok Min, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognit.*, vol. 108, p. 107561, 2020, doi: https://doi.org/10.1016/j.patcog.2020.107561.

[2] Yanjinlkham Myagmar-Ochir, Wooseong Kim, "A Survey of Video Surveillance Systems in Smart City," *Electronics*, vol. 12, no. 17, p. 3567, 2023, doi: https://doi.org/10.3390/electronics12173567.

[3] N. Mumtaz *et al.*, "An overview of violence detection techniques: current challenges and future directions," *Artif. Intell. Rev. 2022 565*, vol. 56, no. 5, pp. 4641–4666, Oct. 2022, doi: 10.1007/S10462-022-10285-3.

[4] Fath U.Min Ullah, Amin Ullah, "Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network," *Sensors*, vol. 19, no. 11, p. 2472, 2019, doi: https://doi.org/10.3390/s19112472.

[5] Simone Accattoli, Paolo Sernani, "Violence Detection in Videos by Combining 3D

Convolutional Neural Networks and Support Vector Machines," *Appl. Artif. Intell.*, vol. 34, no. 4, 2020, [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/08839514.2020.1723876

[6] Romas Vijeikis, Vidas Raudonis, "Efficient Violence Detection in Surveillance," *Sensors*, vol. 22, no. 6, p. 2216, 2022, doi: https://doi.org/10.3390/s22062216.

[7] Guillermo Garcia-Cobo, Juan C. SanMiguel, "Human skeletons and change detection for efficient violence detection in surveillance videos," *Comput. Vis. Image Underst.*, vol. 233, p. 103739, 2023, doi: https://doi.org/10.1016/j.cviu.2023.103739.

[8] Fernando J. Rendón-Segador, Juan A. Álvarez-García, "CrimeNet: Neural Structured Learning using Vision Transformer for violence detection," *Neural Networks*, vol. 161, pp. 318–329, 2023, doi: https://doi.org/10.1016/j.neunet.2023.01.048.

[9] A. Datta, M. Shah, and N. Da Vitoria Lobo, "Person-on-person violence detection in video data," *Proc. - Int. Conf. Pattern Recognit.*, vol. 16, no. 1, pp. 433–438, 2002, doi: 10.1109/ICPR.2002.1044748.

[10] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "Violence Detection in Video Using Computer Vision Techniques," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6855 LNCS, no. PART 2, pp. 332–339, 2011, doi: 10.1007/978-3-642-23678-5_39.

[11] P. Bilinski and F. Bremond, "Human violence recognition and detection in surveillance videos," *2016 13th IEEE Int. Conf. Adv. Video Signal Based Surveillance, AVSS 2016*, pp. 30–36, Nov. 2016, doi: 10.1109/AVSS.2016.7738019.

[12] Muhammad Shoaib, Nasir Sayed, "A Deep Learning Based System for the Detection of Human Violence in Video Data," *Trait. du signal*, vol. 38, no. 6, p. 12, 2021, [Online]. Available: https://www.iieta.org/journals/ts/paper/10.18280/ts.380606

[13] V. Akula and I. Kavati, "Human Violence Detection in Videos Using Key Frame Identification and 3D CNN with Convolutional Block Attention Module," *Circuits, Syst. Signal Process. 2024 4312*, vol. 43, no. 12, pp. 7924–7950, Aug. 2024, doi: 10.1007/S00034-024-02824-W.

[14] Naz Dündar, Ali Seydi Keçeli, "A shallow 3D convolutional neural network for violence detection in videos," *Egypt. Informatics J.*, vol. 26, p. 100455, 2024, doi: https://doi.org/10.1016/j.eij.2024.100455.

[15] Marwa Qaraqe, Yin David Yang, Elizabeth B Varghese, Emrah Basaran & Almiqdad Elzein, "Crowd behavior detection: leveraging video swin transformer for crowd size and violence level analysis," *Appl. Intell.*, vol. 54, pp. 10709–10730, 2024, [Online]. Available: https://link.springer.com/article/10.1007/s10489-024-05775-6

[16] Vesal Khean, Chomyong Kim, "Human Interaction Recognition in Surveillance Videos Using Hybrid Deep Learning and Machine Learning Models," *Comput. Mater. Contin.*, vol. 81, no. 1, pp. 773–787, 2024, doi: https://doi.org/10.32604/cmc.2024.056767.

[17] J. Mahmoodi and H. Nezamabadi-pour, "A spatio-temporal model for violence detection based on spatial and temporal attention modules and 2D CNNs," *Pattern Anal. Appl. 2024 272*, vol. 27, no. 2, pp. 46-, Apr. 2024, doi: 10.1007/S10044-024-01265-0.

[18] P. Zhang, L. Dong, X. Zhao, W. Lei, and W. Zhang, "An end-to-end framework for real-time violent behavior detection based on 2D CNNs," *J. Real-Time Image Process. 2024 212*, vol. 21, no. 2, pp. 57-, Mar. 2024, doi: 10.1007/S11554-024-01443-7.

[19] V. A. M. Chidambaram and K. P. Chandrasekaran, "F3DNN-Net: behaviours violence detection via fine-tuned fused feature based deep neural network from surveillance video," *Signal, Image Video Process. 2024 1811*, vol. 18, no. 11, pp. 7655–7669, Aug. 2024, doi: 10.1007/S11760-024-03418-4.

[20] Gurmeet Kaur, Sarbjeet Singh, "Revisiting vision-based violence detection in videos: A critical analysis," *Neurocomputing*, vol. 597, p. 128113, 2024, doi: https://doi.org/10.1016/j.neucom.2024.128113.

[21] Pablo Negre, Ricardo S. Alonso, "Literature Review of Deep-Learning-Based Detection of Violence in Video," *Sensors*, vol. 24, no. 12, p. 4016, 2024, doi: https://doi.org/10.3390/s24124016.

[22] Fernando J. Rendón-Segador, Juan A. Álvarez-García, "Transformer and Adaptive Threshold Sliding Window for Improving Violence Detection in Videos," *Sensors*, vol. 24, no. 16, p. 5429, 2024, doi: https://doi.org/10.3390/s24165429.

[23] Abu Bakar Siddique Mahi, Farhana Sultana Eshita, "VID: A comprehensive dataset for violence detection in various contexts," *Data Br.*, vol. 57, p. 110875, 2024.

[24] Damith Chamalke Senadeera, Xiaoyun Yang, Dimitrios Kollias, Gregory Slabaugh, "CUE-Net: Violence Detection Video Analytics with Spatial Cropping, Enhanced UniformerV2 and Modified Efficient Additive Attention," *arXiv:2404.18952*, 2024, [Online]. Available: https://arxiv.org/abs/2404.18952

[25] F. Cao, Y. Miao, and W. Zhang, "Implementation and Application of Violence Detection System Based on Multi-head Attention and LSTM," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 14868 LNCS, pp. 77–88, 2024, doi: 10.1007/978-981-97-5600-1_7.

[26] N. Mumtaz, N. Ejaz, I. Rida, M. A. Khan, and M. Y. Lee, "Towards Real-world Violence Recognition via Efficient Deep Features and Sequential Patterns Analysis," *Mob. Networks Appl.*, vol. 29, no. 4, pp. 1326–1335, Aug. 2024, doi: 10.1007/S11036-024-02319-7/METRICS.

[27] L. Hsairi, S. M. Alosaimi, and G. A. Alharaz, "Violence Detection Using Deep Learning," *Arab. J. Sci. Eng. 2024 5015*, vol. 50, no. 15, pp. 11669–11679, Sep. 2024, doi: 10.1007/S13369-024-09536-Y.

[28] M. Evany Anne, M. Brindha, and N. Sivakumaran, "Advancing intelligent surveillance: A comprehensive hybrid deep learning framework for anomaly detection, violence recognition, and person re-identification," *Multimed. Tools Appl. 2025 8438*, vol. 84, no. 38, pp. 46863–46909, Jul. 2025, doi: 10.1007/S11042-025-21005-8.

[29] N. Tran, H. Nguyen, D. Ly, K. Ngo, and H. D. Nguyen, "Advancing Violence Detection with Graph-Based Skeleton Motion Analysis," *SN Comput. Sci. 2025 66*, vol. 6, no. 6, pp. 595-, Jun. 2025, doi: 10.1007/S42979-025-04118-7.

[30] Hong Huang, Qingping Jiang, "IDG-ViolenceNet: A Video Violence Detection Model Integrating Identity-Aware Graphs and 3D-CNN," *Sensors*, vol. 25, no. 20, 2025, doi: https://doi.org/10.3390/s25206272.

[31] Chenghao Li, Gang Liang, "MEN-VVDF: Multipath excitation network-based video violence detection framework focusing on human activity in keyframes," *J. Vis. Commun. Image Represent.*, vol. 112, p. 104573, 2025, doi: https://doi.org/10.1016/j.jvcir.2025.104573.